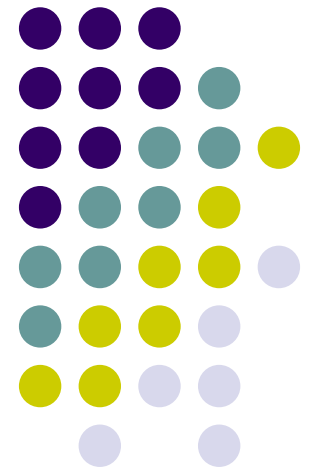


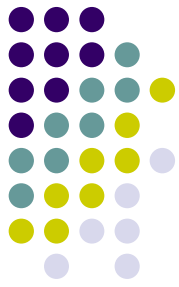
Web Data Management Systeme

Seminar: Web-Qualitätsmanagement

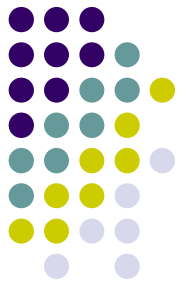
Arne Frenkel



Agenda



- Einführung
- Suchsysteme
 - Suchmaschinen & Meta-Suchmaschinen
 - W3QS
 - WebSQL
 - WebLog
- Information Integration Systems
 - Ariadne
 - TSIMMIS
- Zusammenfassung






Einführung

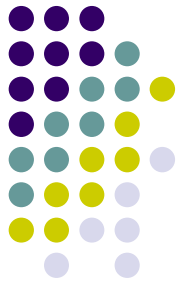
- WWW bietet gute Möglichkeit, Informationen zu verbreiten
- Um Informationen zu finden, werden Konzepte zum Suchen benötigt
- Dabei müssen große Mengen von Daten bewältigt werden, die im gesamten WWW verteilt sein können



Suchmaschinen

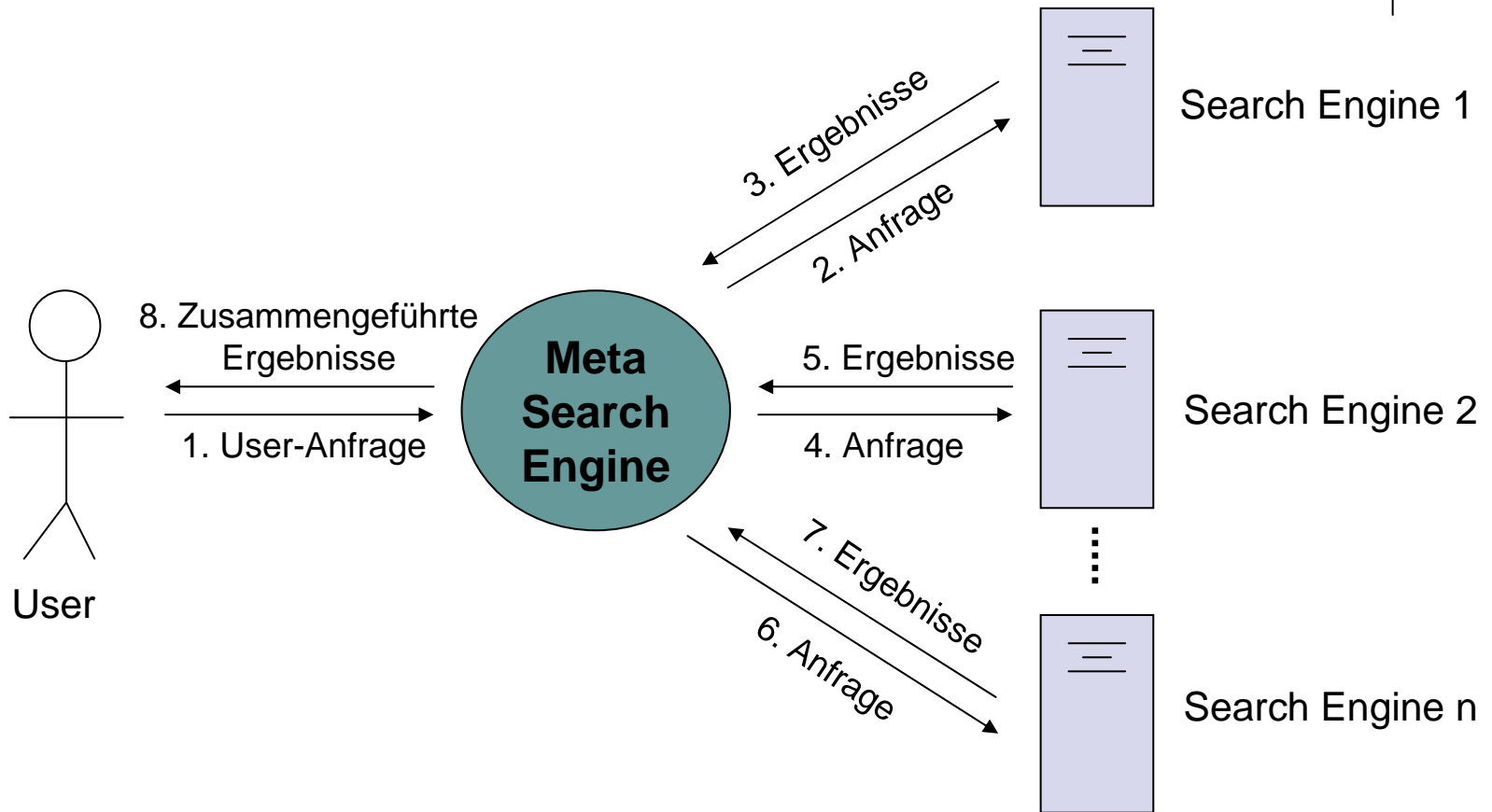
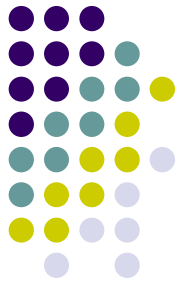
- z.B.  ,  , 
- Unterstützen:
 - Stichwort-Suche
 - Teilweise Suche im Usenet, nach Bildern oder Sound
 - Verknüpfte Suche mit AND, OR, +/-
 - Suchen nach kompletten Phrasen („...“)

Suchmaschinen



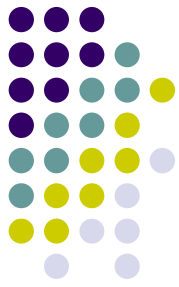
- Nachteile
 - Keine Suche nach Metadaten
 - Keine Suche nach Links
 - Keine Suche nach strukturellen Eigenschaften
 - Nur Text-Suche möglich
 - Mehrfach gefundene Dokumente
 - Keine Suche nach dynamischen Inhalten

Meta-Suchmaschine



W3QS

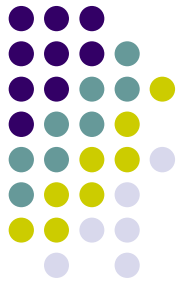
(WWW Query System)



- Technion (Israel Institute of Technology)
- SQL-Ähnlich
- Unterstützt:
 - Inhaltliche und strukturelle Anfragen
 - Eigene Programme können eingebunden werden
 - Online Formulare automatisch ausfüllen

W3QS

(WWW Query System)



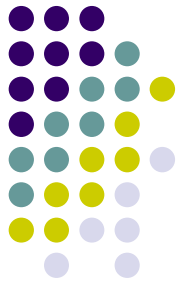
- WWW wird als Knoten (Webseiten) und Kanten (Links) gesehen
- Es wird ein Startknoten festgelegt!

```
SELECT n2
FROM n1, l1, n2
WHERE n1.url = „http://www.uni-
                magdeburg.de“
AND n2.title LIKE „%Informatik%“
```

- Gesucht werden alle Seiten, die über einen Link von n_1 zu erreichen sind und den Titel Informatik enthalten

W3QS

(WWW Query System)

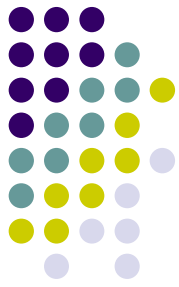


Pfad von unbestimmter Länge

```
SELECT
FROM n1, l1, (n2, l2), l3, n3
WHERE n1.url = „http://www.uni-magdeburg.de“
      n3: PERLCOND , (n3.title.content =
                    /Informatik/i) ` ;
Using ISEARCHd -d 5 -l 1000
```

-d bestimmt die maximale Länge des Pfads (n₂, l₂)
-l bestimmt die maximale Anzahl von HTTP-Requests

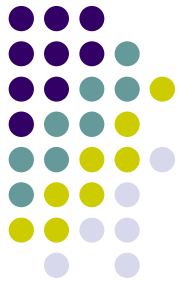
PERL-Programm kann verschiedene Dateiformate analysieren (z.B. HTML, Latex, Postscript)



WebSQL

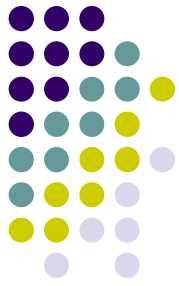
- Universität von Toronto
- SQL-Ähnlich
- Unterstützt:
 - Strukturelle Anfragen
- Annahme:
 - Relationen für Dokumente und Links
 - Document[url, title, text, type,...]
 - Anchor[base, label, href]

WebSQL



- **Verschiedene Arten von Links**
 - Interior: wenn Ziel und Basis-Dokument gleich sind (\mapsto)
 - Local: wenn sich Ziel und Basis-Dokument auf gleichem Server befinden (\rightarrow)
 - Global: wenn sich Ziel und Basis-Dokument auf verschiedenen Servern befinden (\Rightarrow)

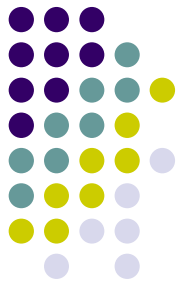
WebSQL



```
SELECT d.url, d.title
FROM Document d
  SUCH THAT „http://www.uni-magdeburg.de“ →|⇒ d
WHERE d.title CONTAINS „Informatik“
```

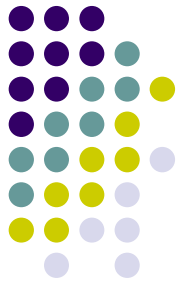
- Ausgegeben werden URL und Titel aller Seiten, die über einen internen oder externen Link von www.uni-magdeburg.de zu erreichen sind und deren Titel Informatik enthält

WebLog



- Concordia Universität
- Logik-basierter Ansatz
- HTML-Tags können zur strukturierten Suche genutzt werden
- Leider wird nur die Suche von einer Seite aus unterstützt

WebLog



- WebLog stellt eine Reihe von Prädikaten zur Verfügung (z.B. `<url> [<attr> -->> <val>]`)

```
informatik_urls.html[title -->'Informatik URLs',  
    hlink -->> L, occurs -->> T]  
  
<-- http://www.uni-magdeburg.de [hlink -->> L],  
    href(L,U), U[title -> T],  
    substring(T, 'Informatik')
```

Information Integration Systems



- Anfragen an verschiedene Online-Quellen
- Um komplexere Antworten zu erlangen
- wichtige Aufgaben:

Strukturierung

- Daten von HTML-Seiten müssen Strukturiert werden
- **Wrapper-Programme** interpretieren eine Anfrage und geben strukturierte Ergebnisse

Kombinieren

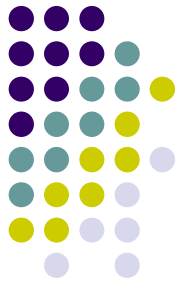
- **Mediatoren** wählen Online Quellen aus und speichern Ergebnisse
- Sie gewinnen zu dem Informationen aus den Ergebnissen

Ariadne



- Erleichtert Erstellen von Wrappern durch Tools
 - Web-Quellen sollen wie DB genutzt werden
 - Stalker inductive-learning system
- Unterstützt Entwurf von eigenen Mediatoren
 - Zum Gewinnen, Abfragen und Integrieren von Daten im Web

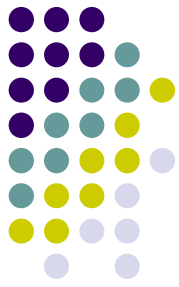
Ariadne



- 2-Phasen Algorithmus
 - 1. Phase: Preprocessing
 - 2. Phase: Planning-by-Rewriting

TSIMMIS

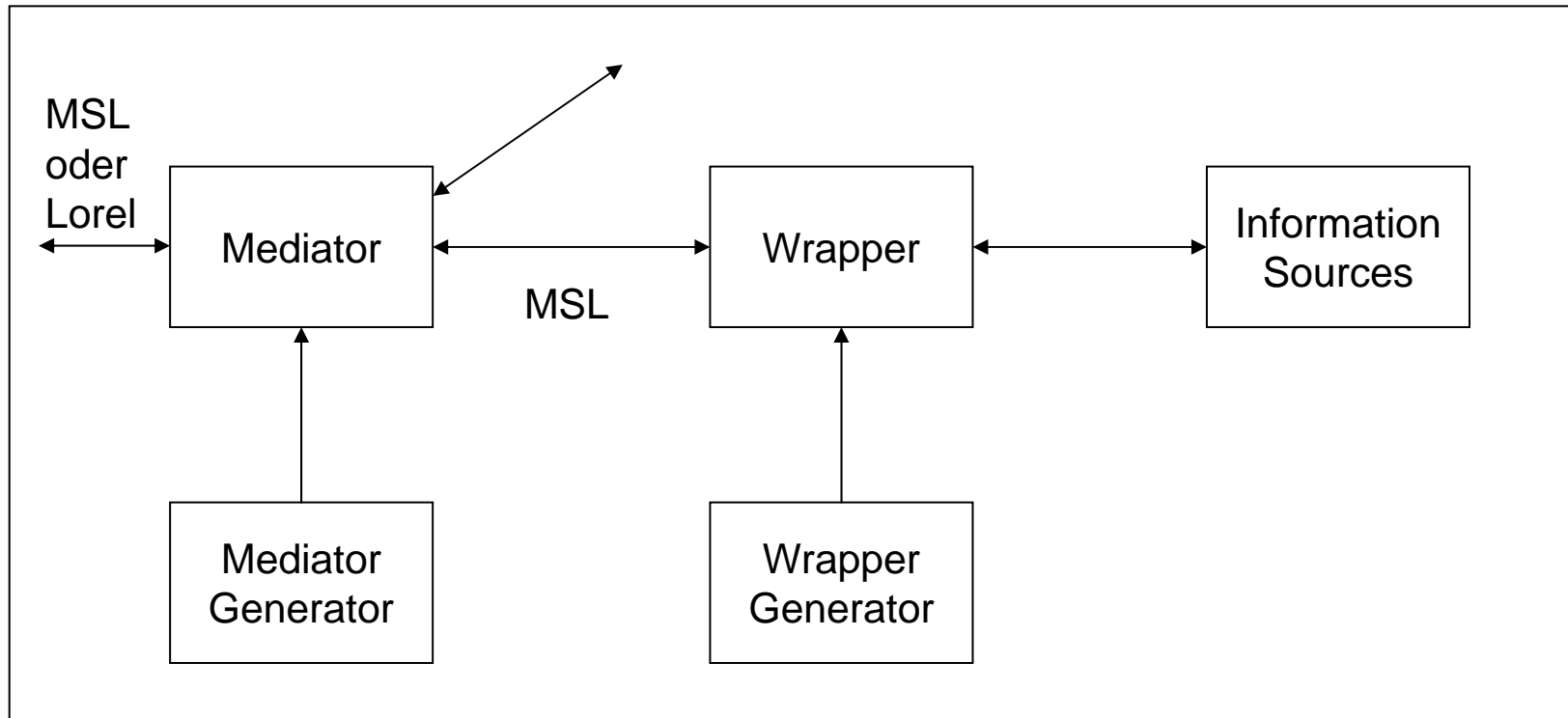
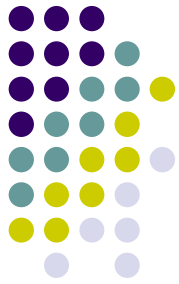
(The Stanford-IBM Manager of Multiple Information Sources)



- Bietet Datenmodell und Anfrage-Sprache
- Verschiedene Informationen über gleiche reale Entitäten können kombiniert werden
- Außerdem bietet es Tools, um Komponenten zum Integrieren von Informationen zu erstellen

TSIMMIS

(The Stanford-IBM Manager of Multiple Information Sources)



Komponenten von TSIMMIS

Zusammenfassung



- Such-Systeme
 - Internet wird als große DB gesehen
 - Meist wird von einem Startpunkt aus eine Suche angestoßen
- Information Integration Systems
 - Informationen von verschiedenen Quellen werden strukturiert und kombiniert